Разведанализ данных

В исследовании мы используем усредненные данные по **50 штатам США с 2010 по 2019 год.**

Чтобы изучить экономические условия штата, мы включили такие характеристики как количество совершенных насильственных преступлений, уровень бедности и безработицы, а также доход и реальный ВВП населения.

Данные были собраны с разных ресурсов, так что требовали некоторых преобразований для дальнейшего анализа.

Предварительный просмотр данных:

Show 10 **entries** **Search:**

	State	Vcrime	PR	UR	PIPC	RGDP
1	Alabama	462.25	0.1805	0.0657017690565541	38436.5	193018.02
2	Alaska	702.87	0.1084	0.0676240971098351	55765.2	54741.09
3	Arizona	432.71	0.1673	0.0688650774555567	39264.4	286210
4	Arkansas	506.21	0.1813	0.0558507627847626	38907.7	111633.63
5	California	433.69	0.15	0.0764721540817871	53327.5	2335825.91
6	Colorado	336.26	0.1173	0.0527386225249409	50676.4	306193.02

Showing 1 to 6 of 6 entries Previous 1 Next

Для подробного анализа мы также проверили корреляцию между зависимой переменной Vcrime и объясняющими переменными:

```
       Vcrime       PR       UR     PIPC
PR    0.41702
UR    0.46002  0.51298
PIPC -0.17195 -0.74873 -0.19432
RGDP  0.14664  0.09647  0.33579  0.25378
```

Рассмотрим обощающие данные по всем переменным:

Переменная State

```
   Length     Class      Mode
       50 character character
```

Текстовая переменная, обозначающая название штата.

```
 [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
 [5] "California"     "Colorado"       "Connecticut"    "Delaware"
 [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
[13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
[17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
[21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
[25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
[29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
[33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
[37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
[41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
[45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
[49] "Wisconsin"      "Wyoming"
```

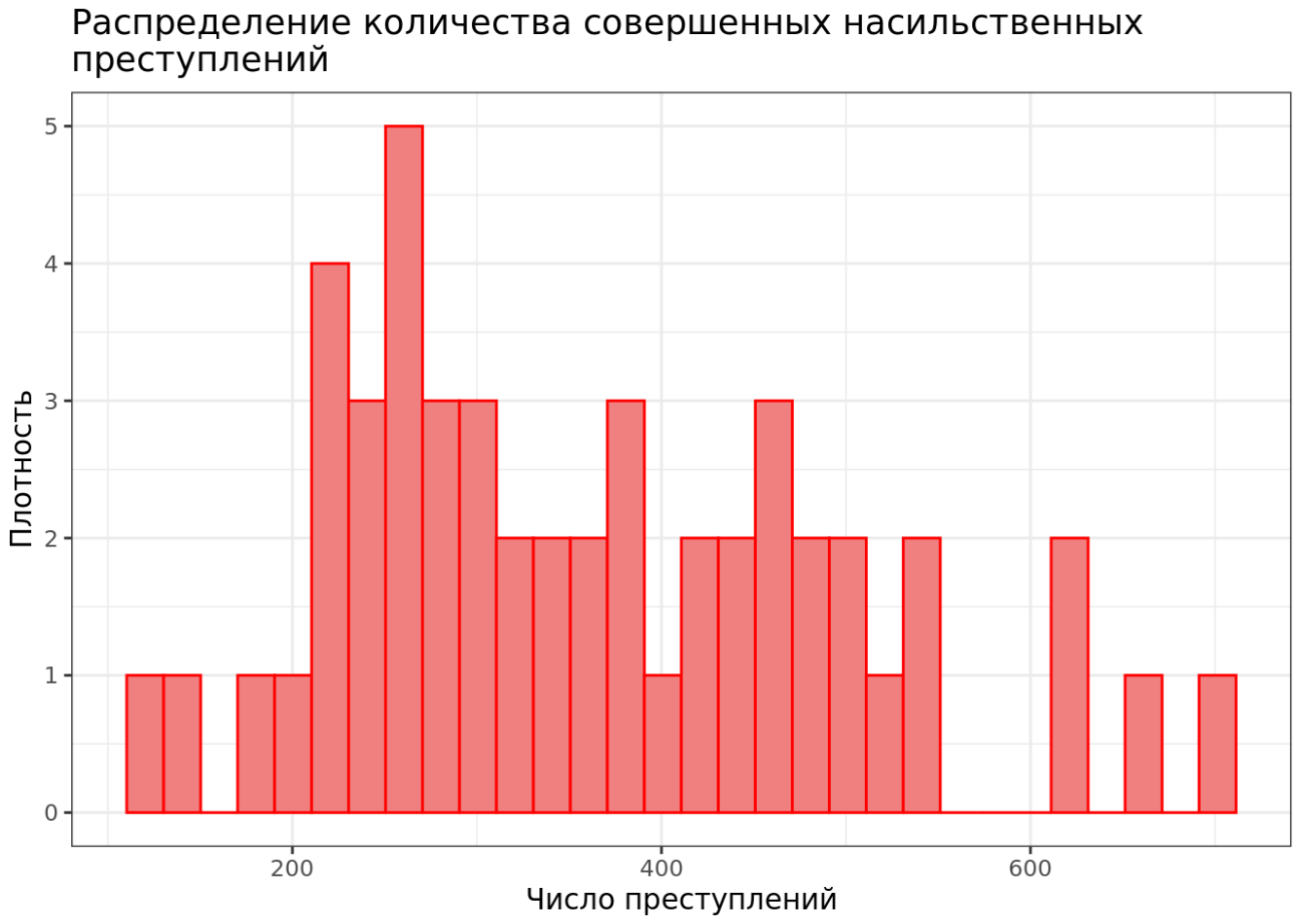
Имеет 50 уникальных значений.

Переменная Vcrime

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  121.7   261.1   337.0   362.9   460.5   702.9
```

Числовая переменная, указывающая количество совершенных насильственных преступлений (violent crimes) на 100 тысяч населения.

Значения колеблются в диапазоне от 122 до 703, где среднее (363) и медианное (337) близки друг к другу.



На графике видно, что количество регистрируемых наблюдений в целом находится на достаточно высоком уровне, что еще раз подтверждает выводы СМИ о существующей проблеме с насильственными преступлениями на территории Соединенных Штатов Америки. Статистические службы также зафиксировали и несколько экстремально высоких значений за все время наблюдения (более 600 преступлений). Подобные выбросы можно объяснить снижением затрат на финансирование правоохранительных органов, усилением расслоения общества, экономическими шоками. Следует заметить, что это могло произойти и вследствие ошибок при фиксации и регистрации насильственных преступлений.

Переменная PR

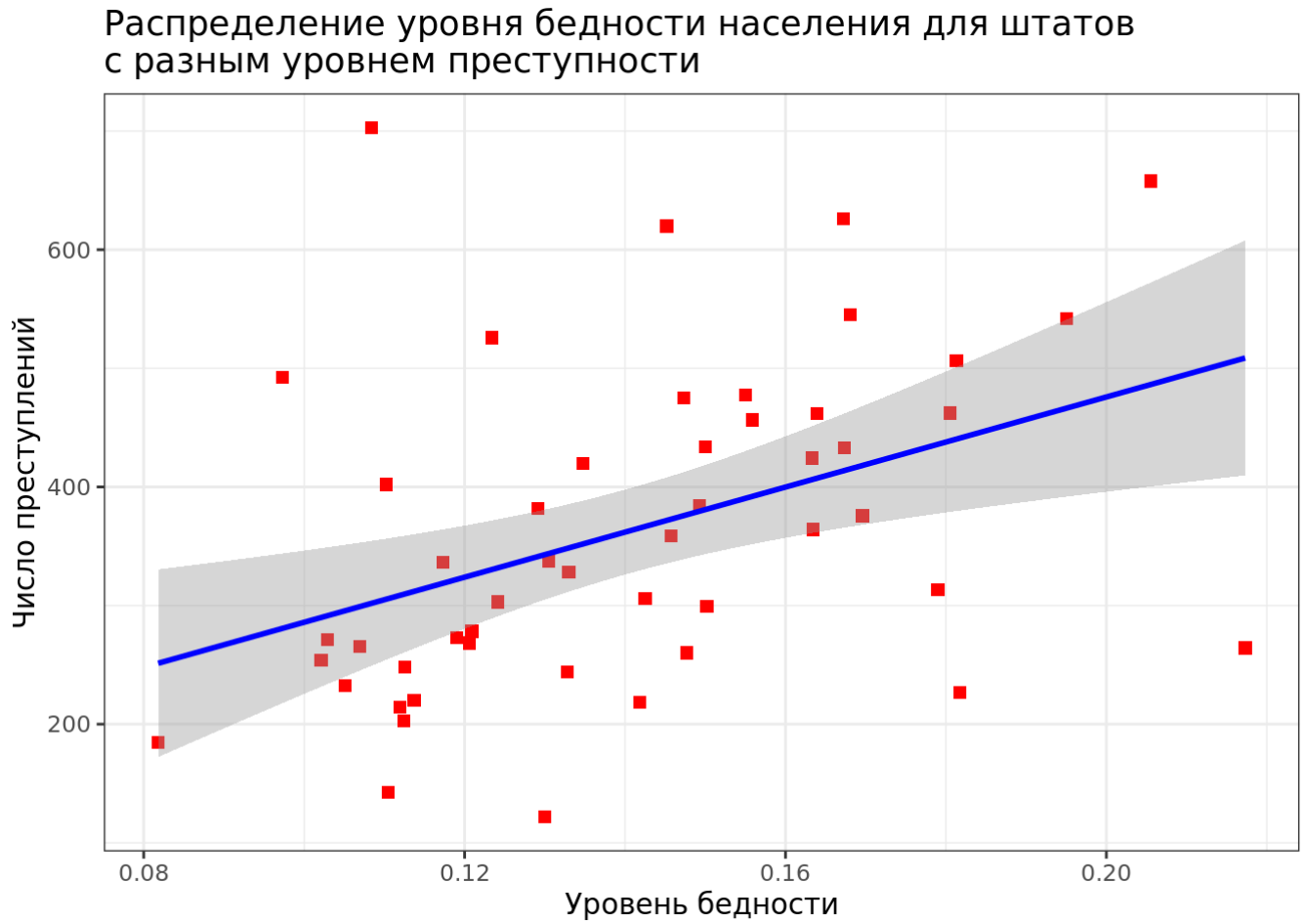
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0818  0.1146  0.1383  0.1405  0.1634  0.2173 
```

Poverty rate - числовая переменная, обозначающая уровень бедности в долях, то есть от 0 до 1.

В нашем случае минимальное значение 0.0818, а максимальное 0.2173.

Коэффициент корреляции с переменной Vcrime равен 0.417, что говорит о несильной положительной взаимосвязи.

На графике рассеивания мы также можем заметить этот тренд:



Дополнительно построим парную регрессию между переменными:

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   95.991     85.818  1.1185 0.268902
PR          1899.409    597.520  3.1788 0.002589 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value равно 0.00259, что говорит о значимости взаимосвязи между этими переменными.

Также проверим распределение данных на гетероскедастичность с помощью теста Бройша-Пагана, так как график был не однозначен, а наличие этого признака может помешать при дальнейшей оценке.

```
    studentized Breusch-Pagan test

data:  lm1
BP = 0.032846, df = 1, p-value = 0.8562
```

В силу большого p-value в данном случае гипотеза об условной гомоскедастичности не отвергается, то есть гетероскедастичности нет.

Что касается данных в целом, почти все штаты показывают неплохие - относительно мировых - результаты по уровню бедности. Тем не менее, есть штаты, которые за десять лет наблюдений все так же имеют достаточно высокую (близко к 0,2 и выше) долю населения, находящегося за чертой бедности. Это может объясняться локальными экономическими, географическими и климатическими особенностями, которые мешают развитию штата.

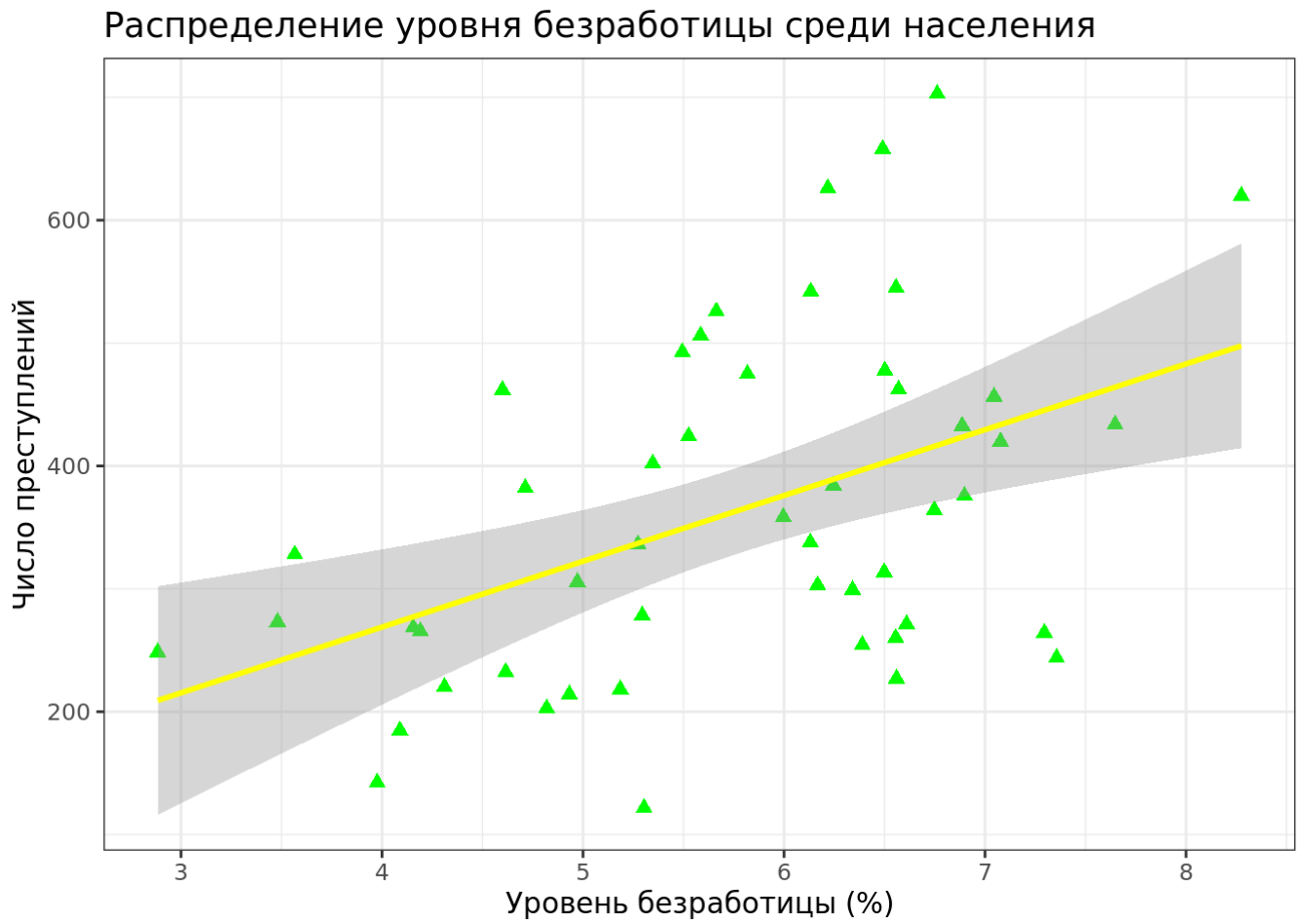
Переменная UR

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02884 0.04942 0.06064 0.05755 0.06559 0.08275
```

Unemployment rate - уровень безработицы в штатах в долях. Значения колеблются от 2,9% до 8,3%.

Коэффициент корреляции с переменной Vcrime равен 0.46, что говорит о слабой положительной взаимосвязи.

Точечный график показывает похожую информацию:



Проверим дополнительно взаимосвязь между переменными в парной регрессии:

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   54.690     87.597  0.6243 0.5353656
UR          5354.747   1491.795  3.5895 0.0007762 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value равно 0.000776 - подтверждаем гипотезу о взаимосвязи.

Также проведем тест Бройша-Пагана, аналогично первому случаю:

```
    studentized Breusch-Pagan test

data:  lm2
BP = 3.2027, df = 1, p-value = 0.07352
```

P-value достаточно велик, чтобы не отвергать гипотезу и утверждать, что гетероскедастичности здесь нет.

Следует также заметить, что уровень безработицы в США в среднем держится на оптимальном уровне, потому что падения ниже 5%, как правило, приводят к снижению общего уровня выпуска и усилению инфляции. Подобные значения также говорят об эффективности работы Федеральной Резервной Системы.

Переменная PIPC

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  35163   41759   45521   46785   50864   67553 
```

Personal income per capita - числовое значение, выражающее доход на душу населения в долларах.

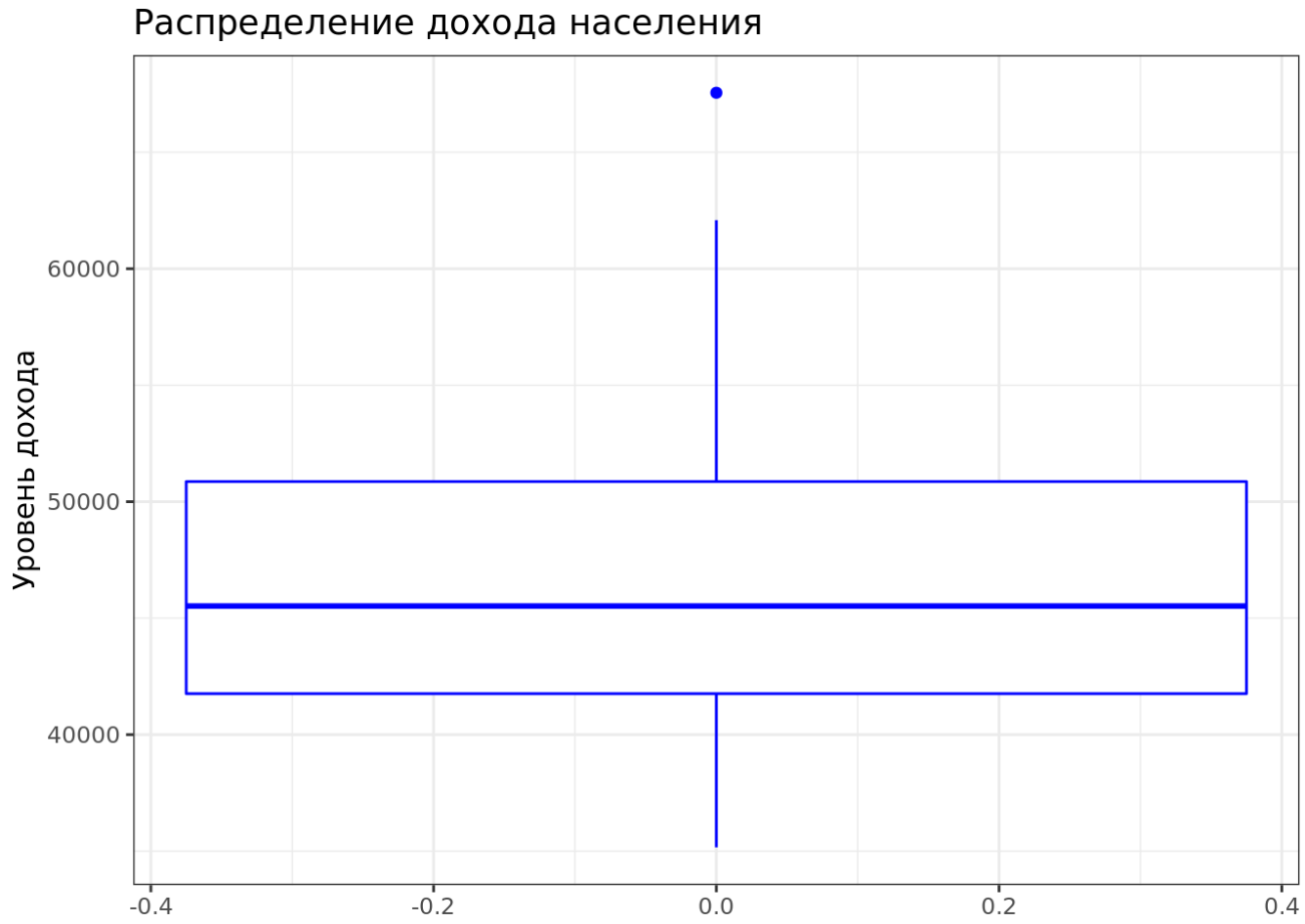
Максимальное значение (67553) почти в два раза больше минимального (35163), хотя среднее и медианное значения - чуть больше 45000. То есть значений, близких к максимальному, довольно мало.

Коэффициент корреляции с Vcrime слишком мал, а значение p-value в парной регрессии слишком велико, чтобы говорить о сильной значимости взаимосвязи переменных:

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept) 518.1785135 129.8798938  3.9897 0.0002252 ***
PIPC         -0.0033200   0.0027453 -1.2093 0.2324580    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рассмотрим график типа boxplot:



Уровень дохода приходится в основном на промежуток от 43 до 51 тысячи долларов на душу населения. Это позволяет сделать вывод о том, что доходы жителей большинства штатов находятся именно в этом отрезке.

Тем не менее присутствует значение, которое выделяется из массы остальных. Проверим штаты с уровнем дохода больше 60000:

Show 10 entries Search:

	State	PIPC
1	Connecticut	67553.1
2	Massachusetts	62083.9

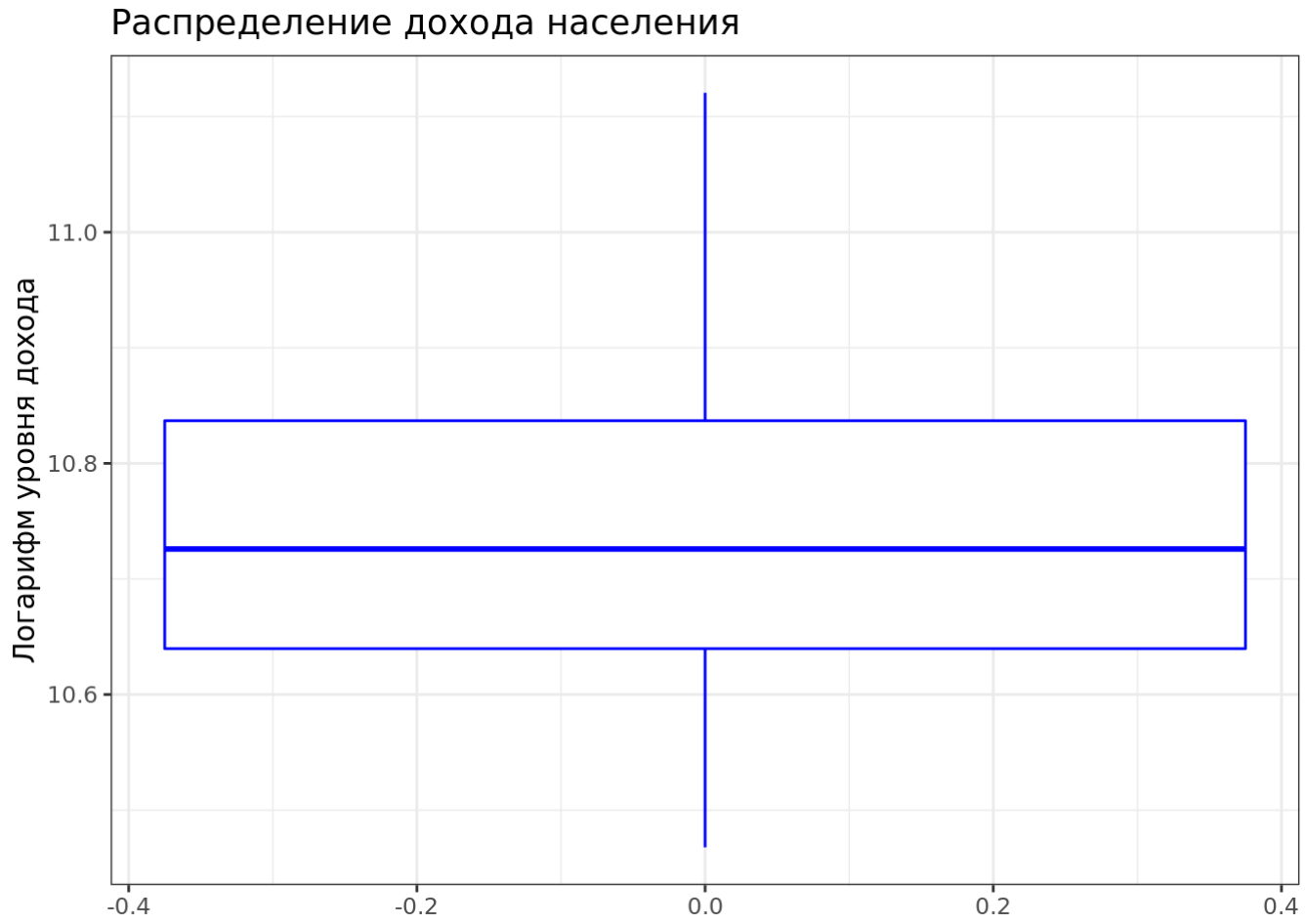
Showing 1 to 2 of 2 entries Previous 1 Next

Выброс в данных - значение уровня дохода в штате Коннектикут. Детально изучив составляющие экономики штата мы сделали вывод, что большую долю в ней составляет доход от военно-промышленных корпораций, так как Коннектикут - главный арсенал страны, а также страховой бизнес - один из наиболее популярных в штате.

Массачусетс - один из мировых центров биотехнологий, искусственного интеллекта и венчурного капитала. Большинство рейтингов ставят данный штат в список лучших для ведения бизнеса. Совокупность этих факторов позволяет ему быть в числе лидеров по показателю дохода на душу населения, хоть это и не является выбросом, судя по графику.

Выбросы можно объяснить спецификой показателя дохода: он фиксирует все доходы человека (зарплата, доходы от аренды, дивиденды, трансферты и другие виды). Это приводит к следующей ситуации: в маленьких штатах показатель PIPC может быть либо слишком высоким, либо слишком низким из-за небывалой урожайности, природной катастрофы или крупного экономического проекта. В то же время дополнительное население (например, студенты колледжей из других штатов) может занижать показатели PIPC.

Скорее всего, видимый выброс в данном случае влияет на корреляцию между переменными, так что попробуем прологарифмировать значение PIPC. Логарифм делает показатели относительными, что позволяет избавиться от выбросов.



Мы видим, что таким образом получилось избавиться от выброса.

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2109.77    1421.32  1.4844   0.1442
logPIPC      -162.62     132.30 -1.2292   0.2250
```

```
          Vcrime       PR       UR     PIPC     RGDP
PR       0.41702
UR       0.46002  0.51298
PIPC    -0.17195 -0.74873 -0.19432
RGDP     0.14664  0.09647  0.33579  0.25378
logPIPC -0.17469 -0.77363 -0.22181  0.99609  0.26278
```

Однако увеличить значимость при этом не удалось. Показатель p-value сдвинулся в меньшую сторону совсем мало, а показатель корреляции в свою очерель поменял знак, но по модулю чуть увеличился. Поэтому в дальнейшем будем использовать именно логарифм дохода населения.

Проверим данные на гетероскедастичность:

```
    studentized Breusch-Pagan test

data:  lm3_
BP = 0.21666, df = 1, p-value = 0.6416
```

Для новых данных гетероскедастичность отсутвует из-за большого p-value.

В рамках показателя Personal Income per Capita наблюдается положительная тенденция: усреднённые за 10 лет данные показывают, что доход на душу населения находится на достаточно высоком уровне почти во все штатах, что свидетельствует о приемлемом состоянии экономики. Кроме того, следует отметить, что подобные показатели достигаются, в том числе, благодаря западной контрактной системе с фиксированными почасовыми ставками.

Переменная RGDP

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  29180   78561  191783  339150  442790 2335826 
```

Real GDP - числовое значение реального ВВП штата. За 10 лет он колеблется от 29180 до 2335826.

Аналогично предыдущему результату нельзя говорить о значимости переменной из-за низкого коэффициента корреляции и большого p-value в парной регрессии.

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.4677e+02 2.4918e+01  13.916   <2e-16 ***
RGDP        4.7435e-05 4.6186e-05   1.027   0.3096    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Максимальное значение = 2335826, а значение 3го квартиля = 442790. Это говорит об очень сильных выбросах. Посмотрим график boxplot:

Тут видны выбросы в лице предположительно 3 штатов. Проверим:

Show 10 entries Search:

	State	RGDP
1	California	2335825.91
2	Texas	1550373.9
3	New York	1371720.34
4	Florida	849951.18
5	Illinois	741158.3

Showing 1 to 5 of 5 entries Previous 1 Next

Действительно, мы видим, что 3 штата - Калифорния, Техас и Нью-Йорк - отличаются от других штатов тем, что средний реальный ВВП с 2010 по 2019 год был выше 1300000, а у остальных штатов значительно ниже.

Нетрудно объяснить подобное отклонение реального ВВП:

- Калифорния - крупнейшая экономика в рамках Соединенных Штатов Америки, пятая по показателям ВВП даже среди стран. Одними из драйверов подобного успеха являются численность населения и диверсификация экономики (финансовый сектор и недвижимость, информационная отрасль, производственный сектор вносят наибольший вклад).
- Экономика Техаса - вторая по размерам после калифорнийской. Данный штат стремительно развивается уже многие годы, и это однозначно отражается на его показателях: первое место по темпам роста ВВП, первое место по экспорту в США. Экономика Техаса во многом зависит от торговли и обилия крупного бизнеса.
- Стоит отметить и экономику штата Нью-Йорк - третью по размерам в США и двенадцатую - по всему миру. Штат является центром всех финансовых операций: подтверждением тому служит Нью-Йоркская фондовая биржа. Кроме того, Нью-Йорк - крупный морской порт.

Однако эти выбросы снова могут отрицательно повлиять на наши будущие оценки, поэтому логарифмируем переменную.

От выбросов избавились. Проверим числовые показатели:

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0817   227.7065  0.0311   0.9753
logRGDP      29.1599    18.5975  1.5679   0.1235
```

```
          Vcrime       PR       UR     PIPC     RGDP  logPIPC
PR       0.41702
UR       0.46002  0.51298
PIPC    -0.17195 -0.74873 -0.19432
RGDP     0.14664  0.09647  0.33579  0.25378
logPIPC -0.17469 -0.77363 -0.22181  0.99609  0.26278
logRGDP  0.22073  0.13365  0.42050  0.17917  0.82532  0.18031
```

Нельзя сказать, что данные значительно изменились, но все же показатели стали ближе к тем, что гарантируют значимость. Именно поэтому мы будем использовать логарифм реального ВВП в дальнейшей модели.

Также проверим данные на гетероскедастичность:

```
    studentized Breusch-Pagan test

data:  lm4_
BP = 4.5768, df = 1, p-value = 0.03241
```

Из-за маленького p-value гипотеза о гомоскедастичности отвергается - данные гетероскедастичны, поэтому в оценке будем считать робастные ошибки.

В рамках Real GDP наблюдается тенденция к росту показателя в зависимости от размера и численности населения штата, его ранга бизнес-среды, скопления высокотехнологичных производств и размерности рынка венчурного капитала.

Общий вывод

Нам удалось собрать датасет с усредненными данными за обширный период времени, на основе которого мы сможем проверить нашу гипотезу о наличии зависимости между экономическими условиями штата и количеством насильственных преступлений в нем. В рамках разведанализа нам удалось провести необходимую чистку и изменение данных и убедиться в правильности выбора переменных в рамках нашего исследования.